Improving the quality of development assistance

what role for qualitative methods in randomized experiments?

Prowse, Martin; Camfield, Laura

Published in:
Progress in Development Studies

Publication date:
2013

Document version
Peer reviewed version

Citation for published version (APA):
Prowse, M., & Camfield, L. (2013). Improving the quality of development assistance: what role for qualitative methods in randomized experiments? *Progress in Development Studies*, *13*(1), 51-61. http://pdj.sagepub.com/content/13/1/51.abstract

Progress in Development Studies 13, 1 (2013) pp. 51–61

Improving the quality of development assistance: What role for qualitative methods in randomized experiments?

Martin Prowse

Department of Geography and Geology, University of Copenhagen, Øster Voldgade 10, 1350 Copenhagen, Denmark

Laura Camfield

School of Development Studies, University of East Anglia, UK

Abstract: While randomized experiments can be valuable tools in evaluating aid effectiveness, research designs limit the role of qualitative methods to 'field visits' or description of contexts. This article suggests expanding the role of qualitative methods and highlights their advantages and limitations relative to survey methods. It reviews a range of qualitative methods and suggests that life histories are compatible with the internal and external validity criteria of randomized experiments. It illustrates this with a case study of their proposed use in an evaluation of the promotion of *Jatropha curcas*, a second-generation biofuel, in Malawi.

Key words: randomized experiments, qualitative methods, methodology

I Introduction

The effectiveness of development assistance has come under close scrutiny in recent years with the terms of debate shifting from the quantity of aid towards improving quality, particularly through improving the evaluation of aid's impact. For example, the internal focus of many donor evaluations on policy and strategy has obscured the impact of aid on the well-being of recipients, which has made it difficult for agencies and governments to attribute improvements in well-being to specific policy interventions. Some argue that this has contributed to a lack of evidence and consensus around that simplest of questions: what works? (For example, see Banerjee *et al.*, 2007; Savedoff *et al.*, 2006.)

Thus, there has been an upsurge in interest in impact evaluation methodologies (see, for example, a recent DFID-commissioned paper

 10.1177/146499341201300104

on impact evaluation methods by Garbarino and Holland, 2009). It is not our purpose to compare and contrast different approaches to evaluation, and certainly not to try and posit a hierarchy of techniques (which, in any case, must surely depend, *inter alia*, on the research questions in hand, available resources and expertise of the investigators). Instead, we focus on one type of impact evaluation – namely, randomized control trials – to assess the extent to which qualitative research methods could play a central role in an experimental design.

This question is important because many influential contributions to the debate on impact evaluation barely mention qualitative methods (for example, Savedoff *et al.*, 2006). This is not surprising as randomized experiments within international development have been promoted by micro-econometricians who mainly favour survey measurement (although not always – see Chattopadhyay and Duflo, 2004). However, researchers associated with Network of Networks for Impact Evaluation (NONIE) and the International Initiative for Impact Evaluation (3IE) have long argued that qualitative methods should play a complementary, if secondary role to rigorous quantitative methods (White, 2008). Moreover, Karlan (2009: 2) has rightly stated that 'the decision about what to measure and how to measure it, i.e., through qualitative or participatory methods versus quantitative survey or administrative data methods, is independent of the decision about whether to conduct a randomized trial', and outlines further studies that utilize non-quantitative methods. Whilst the acknowledgment that qualitative methods can be utilised within a randomized experiment is to be welcomed, Karlan (2009) says little about how this should be done, even though the inclusion of qualitative methods is common within the related field of social policy (see Molloy *et al.*, 2002).

This paper assesses the extent to which different qualitative research methods could be used as the primary measurement tool within a randomized design. It outlines the advantages of qualitative methods relative to the survey method and assesses the extent to which they can adhere to the basic characteristics of randomized design. The paper argues that two qualitative methods – life history interviews and semi-structured interviews – appear suitable, and focuses on the former, illustrating their value using a case study of their proposed use in an evaluation of the promotion of *jatropha curcas*, a second-generation biofuel, in Malawi.

II What are randomized experiments?

Randomized experiments are designed and structured to answer a counterfactual question: how would participants' welfare have altered if the intervention had not taken place? They have three main characteristics: first, they focus on the impact of an intervention on welfare/well-being outcomes of participants; second, they use counterfactual analysis; and third, they necessitate substantial primary research. They use randomization because this overcomes important limitations in many non-experimental studies such as selection bias and attribution issues**.** In other words, participants in any program are unlikely to be a random sample of the population as a whole (as programs are often 'targeted' at specific groups, or particular social strata self-select). Randomizing who receives an intervention overcomes selection bias by trying to ensure that both the known and unknown characteristics of control and treatment groups are similar (although many practitioners recognise that attaining this level of comparability across participant groups within a community or society is not straightforward due to the existence of 'unobserved' or 'essential' heterogeneity – Heckman *et al.*, 2006 in Ravallion, 2009). Random assignment of who receives an intervention (and who is included in the counterfactual comparison group) allows evaluators to attribute significant change to the intervention in question.

Randomized experiments can be assessed according to the extent to which they adhere to internal and external validity criteria. Internal validity allows the attribution of 'change' to the intervention in question. External validity allows findings to be extrapolated to a wider population (although this may not be possible when the treatment group has specific characteristics, for example, extreme poverty, evidence of child malnutrition). The two criteria are closely related as the control required for absolute internal validity may compromise the ability of findings to be extrapolated to a wider population (Deaton, 2009; Rodrik, 2009).

III The strengths and shortcomings of randomized experiments

Randomized experiments are a powerful tool which can identify the effects of a specific (series of) intervention(s), including components offered to different treatment groups. Their results are easy to convey and often resonate with policymakers and funding agencies, providing a basis for cost–benefit analysis. Additionally, experiments can create a long-term relationship between evaluators (which until now have mainly been econometricians and their research students) and implementing agencies (such as donors or NGOs). This increases the evidence basis for the implementing agency's work and/or supports scaling up by national governments. But just as it is important to be open and realistic about the strengths of randomized experiments, we also have to be explicit and clear about their shortcomings (which until recently have not been discussed with enough candour). Below, we briefly summarize limitations to randomized experiments within the research design itself, which have been acknowledged by practitioners, before addressing some broader issues.

Six limitations that affect internal validity are first, 'attrition from samples', possibly as a result of the intervention or evaluation. This is shared by all types of longitudinal research, and can be partly overcome by tracking people if they move or if the household splits (although this is inevitably costly). Second, the 'merging of treatment and control groups' where a control group forces itself into the treatment group, perhaps due to local or institutional politics. There can also be 'spillover effects' between treatment and control groups such as when an agricultural intervention also increases labour demand in neighbouring communities. While leakage can be mitigated through randomization procedures – for example, increasing the geographical distance over which control and treatment are selected—increasing the distance between groups might also reduce their geographical similarity. Fourth, 'implementing agencies may not comply', for example, by failing to ensure the separation of treatment and control groups. Fifth, there may be 'limited attention to sub groups'. The conventional output from an experiment is the average treatment effect on the treatment (ATT), and thus sub groups are often not reported. This can obscure the losses incurred by some participants. For example, Deaton (2009: 29) states that 'the trial might reveal an average positive effect although nearly all of the population is hurt with a few receiving very large benefits'. In this respect, randomized experiments which solely focus on the ATT are implicitly based on the utilitarian notion of improving aggregate expected utility, in other words, whether an intervention will achieve the greatest good for the greatest number.

Such a perspective conflicts with justice- and rights-based approaches to development which are concerned with the poorest members of societies and ensuring that no individual should fall below minimum thresholds. Deaton (2009: 29) cautions that 'much of the disagreement about development policy is driven by differences of this kind'. And sixth, there may be 'strong moral and ethical concerns' against using portions of a population as a control group. For example, the provision of basic services in health and education is a human right, and withholding such services from a portion of a population as a control group may

be ethically unacceptable and may cause avoidable harm. Proponents of randomized experiments suggest this shortcoming can often be avoided by employing a 'pipeline approach' using communities or households that have been selected for project but not yet treated as the comparison group (thus avoiding selection bias). However, withholding resources from poor people who live in risky environments so that they can constitute a 'control' group can create avoidable harm. Evaluators need to ensure that withholding treatment will not contribute to individuals falling below a minimum threshold that might have a lasting effect on their wellbeing. Engaging with participants using qualitative methods highlights the range of risks they face and enables evaluators to face ethical issues with the seriousness and sincerity they deserve.

External validity is affected by four further limitations. The first of these is 'the influence of context on the intervention'. For example, Deaton (2009: 43) warns that 'an educational protocol that was successful when randomized across villages in India holds many things constant that would not be constant if the program were transported to Guatemala or Vietnam' (also see Woolcock, 2009). This is due to both the influence of the 'socio-cultural and physical environment on the intervention' and changes that take place in the 'implementing institution' when projects are scaled up. For example, Woolcock (2009: 8) highlights the example of the Kecamatan Development Project, Indonesia, which became more successful on scaling up as it learnt from its experiences and was able to attract better quality staff. As Deaton (2009: 44) also notes, 'small development projects that help a few villagers or a few villages may not attract the attention of corrupt public officials […] yet they would do so as soon as any attempt were made to scale up'. The second consideration is that 'interventions can cause changes in behaviour' that would not occur if scaled up (for example, increased uptake at a pilot stage due to the novelty of the intervention). Third, the 'evaluation itself can cause the treatment and/or control groups to change behaviour', for example, if people in the control group view themselves as being in competition with the treatment group and so alter their actions. A related concern among evaluators (for example, Adato, 2007) is that randomized experiments can increase social differentiation and even create conflict between beneficiaries and non-beneficiaries. Equally problematic from the point of view of evaluation is where an intervention creates 'equilibrium effects' when it is scaled up. A good example comes from Banerjee and Duflo (2008): An evaluation might find that extra-curricular tuition for lagging students improves employability post-education. However, if this was scaled up at a national level, the extra supply of school leavers who benefited from this tuition would limit each student's chances of getting a job.

Even when internal and external validity issues are fully taken into account, some scholars are sceptical about the extent to which randomized experiments can generate 'gold standard' data. This is an important consideration given the increasing emphasis on evidence-based policy making, partly informed by systematic reviews, which typically treat experimental data as the highest form of evidence. As suggested earlier, there is a 'familiar trade-off between internal and external validity' as the formal methodology puts severe constraints on the assumptions a target population must meet to justify extrapolating a conclusion outwards from the treatment group (Cartwright, 2007: 11). Deaton (2009: 6) concurs that 'the price for this success [in internal validity] is a focus that is too narrow to tell us "what works" in development, to design policy, or to advance scientific knowledge about development processes'.

A related concern has been the types of interventions selected for evaluation through randomized designs. For example, Jones *et al*. (2009) suggest there are significant gaps in the application of counterfactual impact evaluations (encompassing both randomized

experiments and *ex post* quasi-experimental approaches). In particular, they highlight the lack of studies on environmental protection, agriculture and on gender issues. This line of argument reflects the belief that randomized experiments 'can take only a very specialized type of evidence as input and special forms of conclusion as output' (Cartwright, 2007: 12).

There are also further reasons why randomized experiments are good for addressing certain research questions and not others. Experiments require time to ensure that interventions are embedded before the end-line research wave is conducted, and this may conflict with the short-term policy horizons of governments and donors. In addition, whilst randomized experiments are suited to small-scale development projects, they are not suitable for evaluating broad policy changes. For example, public sector reforms or changes to exchange rates or trade regimes are not appropriate due to the difficulty in establishing the counterfactual. White (2007: 7) comments dryly that it is usually 'not possible to randomly place large-scale infrastructure, such as a port or major bridge'. Moreover, we should not forget political concerns: those with vested interests in a program (perhaps local political elites, or even donor or project staff) may have reasons to try and prevent a randomized evaluation (and prefer the status quo where procedures and impacts are opaque). This suggests the need for a holistic approach to evaluations of complex and politically-sensitive social interventions (perhaps drawing on the experiences of evaluators within social policy who have made extensive use of qualitative methods to capture diversity in outcomes and mechanisms, and explain how these mechanisms work).

IV Mixed methods within an experimental design

Randomized experiments have so far been dominated by quantitative methods, almost exclusively based on the survey instrument. For example, it is rare to see skilled and time-intensive methods such as ethnography used as part of a randomized experiment (although embedding anthropologists within institutions conducting randomized experiments would be highly beneficial). The dominance of quantitative methods is hardly surprising: the experimental methodology adheres to positivist principles and is very good at tackling 'what' and 'where' questions (which means it is good at capturing a state or condition). But, by relying only on quantitative methods randomized experiments are often unable to tell us very much about 'how' or 'why' societal change occurs – they often cannot inform us about key transmission mechanisms and therefore how interventions can or cannot be scaled up or transferred to other settings. Adato (2007: 9–10) notes, for example, that survey methods are at a disadvantage when it comes to unpacking the 'black box' of impact due to:

> The necessary brevity of questions and the use of proxies that are often blunt measures; respondents' inability to sufficiently express what they mean in selecting among categorical or continuous variables; the limited ability of enumerators to follow up when more information or clarification is needed; and the difficulty of establishing the rapport and trust needed to maximize truthfulness in replies.

Qualitative methods, on the other hand, are generally able to shed light on 'why' and 'how' questions, are good at capturing processes, and pay greater attention to why certain individuals benefit from an intervention and others do not. Examples of where qualitative methods have been used in impact evaluations in developing countries include Rao and Ibanez (2005, social funds in Jamaica); Adato (2007, Conditional Cash Transfer Schemes (CCTs) in Nicaragua and Turkey); White's study of education reform in Ghana (see White, 2008); and White and Masset's (2007) study of an integrated nutrition project in rural Bangladesh.

These examples illustrate the value of combining qualitative and quantitative methods within studies, but they do not investigate

whether qualitative or participatory methods could be the 'primary' measurement tool within a randomized design. In the next section we review the following qualitative methods, bearing in mind that qualitative methods can also be used to generate quantitative data:

1. Ethnography (or in other words, participant observation over a relatively long timescale).
2. Semi-structured interviews (where the interview is guided by a checklist of pre-determined open-ended and closed questions).
3. Life history interviews (there are numerous forms of biographical methods—here we refer to a structured elicitation of a respondent's life story which includes co-creating a timeline for the respondent to discuss and interpret and the addition of closed questions).
4. Focus group discussions.
5. Task-based group methods, often used as part of 'Participatory Poverty Assessments', such as community mapping and ranking exercises.

Tables 1 and 2 compare the basic characteristics of randomized experiments described above with the five types of qualitative research. Each method is assessed according to the likelihood that they could adhere to the basic characteristics of randomized experiments.

Table 1 suggests that one method appears unsuitable at this point—ethnography—mainly because of its attention to detail and deep immersion in circumscribed locations. It also reflects the inductive nature of ethnography, where research questions emerge from long-term participation and observation in a community and are usually not clearly defined prior to entering the field. This is not to say that ethnography could not run parallel to the main experimental design, or be used in a mixed method design (Adato, 2007), but that the ethos of ethnography (not to mention the practicalities and cost) militate against using this methods as the 'primary' measurement tool. The same argument applies to genuinely participatory research (for example, participatory learning and action, PLA), which tend not to

Table 1 To what extent might qualitative methods adhere to the basic characteristics of randomized experiments?

	Ethnography (Participant Observation)	Semi-structured Interviews	Life History Interviews	Focus Group Discussions	Task-based Group Methods
Ex ante null hypothesis to be disproved					
Specified causal pathway					
Specified main variables					
Sufficient sample size for data saturation					
Randomly select treatment and control groups					
Research waves before and after intervention					
Data analysis					
Potential for use as primary research tool in experimental design					

Source: Authors.

Note: Two categories: likely (light grey) and unlikely (dark grey).

Table 2 To what extent do qualitative methods compromise the internal and external validity of randomized experiments?

		Semi-structured Interviews	Life History Interviews	Focus Group Discussions	Task-based Group Methods
Internal Validity	Attrition				
	Merging of treatment and control groups				
	Spillover effects				
	No institutional compliance				
	No sub groups				
	Moral or ethical concerns				
External Validity	Context – environmental				
	Context – institutional				
	Pilot creates effects				
	Evaluation changes behaviour				
	Equilibrium effects				
Cost					

Source: Authors.

Note: Whether qualitative methods might do better (light grey), the same (dark grey), or worse (black) than the conventional survey method.

have an *ex ante* hypothesis, a predicted causal chain or *ex ante* selection of main variables due to an inductive and iterative approach to generating research questions (not included in Table 1). However, participatory 'methods', such as task-based group approaches, can be used within an experimental design (as illustrated by Chattopadhyay and Duflo, 2004), as such methods are increasing being used to generate statistics (see Barahona and Levy, 2003) not least as part of participatory impact assessment approaches (see, for example, the work of the Feinstein International Centre at Tufts University).

This leaves us with four possible methods: semi-structured interviews; life history interviews; focus group discussions; and task-based group methods. These four methods are now compared in terms of the extent to which they compromise the internal and external validity of a randomized experimental design.

Table 2 suggests that focus group discussions and task-based group methods may do worse than the survey method in terms of spillover effects and the evaluation changing behaviour, due to the open, public nature of these methods. For example, people may be reluctant to admit to receiving benefits from other sources or to not having changed their behaviour in the intended direction. However, there is no reason to suppose that respondents will reveal this information to an official enumerator they have only just met, and it may be that free discussion within a focus group will give them more confidence to speak frankly. Overall, though, we feel that 'collective' methods will probably perform worse than 'individual' methods. Group methods may also be more expensive, due to higher fixed costs per research encounter (although there may be a trade-off in terms of the numbers required for data saturation, especially within a clustered research design).

On the other hand, semi-structured and life history interviews do not appear to compromise the experimental design to any greater extent than the conventional survey methods. After all, a survey is typically based on a participant's responses in a one-on-one interview and the quality of the data depends on the quality of the rapport between the enumerator and the participant. In this respect, it can be argued that the dialogic nature of semi-structured and life history interviews will improve the quality of data generated. For example, the better rapport between the enumerator and the respondent is likely to increase the fidelity of responses. In addition, these methods can reduce the likelihood of attrition from samples by, for example, assuaging respondents' uncertainty. They can also enable an exploration of why individuals might have altered their practices due to the intervention or the structure of the evaluation (allowing greater insight into the external validity of the experiment).

This brings us to one further point regarding external validity. These qualitative methods, by their nature, are also likely to perform better than the survey tool in understanding contextual threats to the experimental design. This is in terms of both the influence of the socio-cultural and physical environment on the intervention, and whether institutions will act differently if the intervention is scaled up. Whilst this clearly has implications for the piloting of measurement tools, as using a qualitative method within the piloting phase could highlight potential threats, it also has implications for using qualitative methods which are amenable to quantification as the 'primary' measurement tool. For example, qualitative methods can help to explicate how aspects of a local environment (whether political, social or physical) might be idiosyncratic, and can capture institutional peculiarities and possible dysfunctionality to a much greater extent than the survey method.

There is also an argument that such qualitative methods are also much more likely to tell us 'why' an intervention succeeds or fails compared to the survey method. For example, Ahmed *et al.*'s study of a conditional cash transfer in Turkey (2006, in Adato, 2007: 22) demonstrated that the reluctance to send daughters to secondary schools went beyond schooling costs as 'secondary schools are often far from home, and transportation options are not trustworthy with respect to [girls'] honour'. So, even though the CCT alleviated the burden of school expenses and prevailing poverty 'where the other factors were strong, the cash could not compensate' (ibid.). Qualitative methods can tell us about the importance of such key transmission mechanisms and societal norms. In sum, using qualitative methods as the primary measurement tool not only adds contextual explanation to the average treatment effect on the treated, but can offer a much richer and more accurate approximation of causal mechanisms than solely using a survey measurement tool.

V An experimental research design using a qualitative method

This penultimate section now discusses which of these two methods might be best suited for experimental designs. In other words, if a funding agency wanted to allocate scarce resources to conduct randomized evaluations using a qualitative method, which method might be first in line? In our opinion, it could well be life history interviews. Why? There are three reasons.

First, the longitudinal focus of a life history interview resonates with the 'before and after' characteristic of experimental designs. Second, a life history interview highlights the importance of social relations and institutions for assessing the intervention in question (birth, childhood, school, marriage, children, employment perhaps). And third, life history interviews allow the generation of quantitative, qualitative and visual data (which can be cross-checked to resolve mismatches and improve data quality – see Davis and Baulch, 2009).

But that is not to say using life history interviews within an experimental design does not have a number of shortcomings. For example, the cost per interview will be higher (due to the greater duration per research encounter, and fewer interviews per day), expanding the resources required for the study, or reducing the power of the findings. The training of researchers will also be more expensive, as few have experience of conducting this form of research method. Using this retrospective dialogic method also raises ethical concerns: asking individuals to recount the trajectory of their life often brings painful memories to the surface (particularly in developing countries where citizens endure much greater levels of risk). Will researchers be able to disengage from respondents in an ethically acceptable manner?

To ground these arguments we now offer a simple research design for a proposed evaluation of a farmers' organisation's programme to promote second-generation biofuel production in Malawi. The design utilizes life history interviews as the primary data generation tool within an experimental methodology. The evaluation will randomly select 88 farming clubs from a key agricultural region who expressed an interest in adopting *jatropha* (whilst random selection of the population is not necessary within randomized experiments, we conduct it here to increase the external validity). The number of clubs and households is determined through assessing the most efficient combination within a clustered design (taking into account an intra-club correlation figure of up to 0.24), utilizing impact variable data from a 2001/02 household survey.

All farming clubs will take part in an initial focus group discussion and all treatment and control clubs will receive extension and training for burley tobacco, groundnuts and soya beans (the conventional export crops in the region) during the agricultural off season. Half the clubs will also be randomly assigned to receive the farmers' organization's package of support to promote *jatropha* production (the extension package includes land preparation, seedling propagation, out-planting practices, field management, harvesting, post-harvest handling, and household-level *jatropha* processing).

After one agricultural season, six households will be randomly selected from each of the 88 clubs, and life history interviews will be conducted with the head of each household. The interview will start by asking broad questions about the history of the household, kinship ties, basic household characteristics and well-being in the household. It will then chart changes in well-being throughout the respondents' life-course, focusing particularly on the period since the household was formed. During this discussion the respondent will be encouraged to complete a visual trajectory of their well-being through time, with key moments of improved well-being and harm investigated. Close to the end of the interview, closed questions regarding income/expenditure, crop production, food purchases in local markets, off-farm and non-farm livelihood strategies, and the intra-household division of labour and income will be discussed. The interview will be conducted in the vernacular, and will be digitally recorded (after asking for the consent of the respondent). At the same time, a complementary life history interview will be conducted with the most senior woman in the household, where applicable (the GPS coordinates of the household will be noted, and a photo will be taken to act as an *aide memoire* when analyzing qualitative data). All households will be compensated for their time and will receive transfers in cash or in kind to ensure *jatropha* adoption does not harm adopting households (the yield and profitability of *jatropha* is extremely uncertain).

The same evening enumerators will listen to the interviews and produce an annotated version of the well-being trajectory. At the same time, enumerators will offer a translation of the key components of the interview which will be fleshed out in full after the completion of the first wave of life histories. Any obvious mismatches between the qualitative, quantita-

tive and visual material within the data from the household head will be resolved through a further visits to the household. After all 528 household heads have been interviewed, the full digital interviews of both household heads and, where applicable, wives will be transcribed and translated in full.

As *Jatropha* yields are always low in the first year increasing to peak yields from the third year onwards, two further waves of interviews will be conducted with respondents, after two years and three years of production, respectively. This will capture the non-linear trajectory of impact on smallholders' food security status.

The second and third waves of the research will utilize the annotated trajectory as the basis for the interview, and the respondent will be asked to amend or add to the well-being chart constructed from the initial interview. The interview will discuss changes in the intervening period by revisiting the kinship ties and basic demographic characteristics. The respondent will then be asked to illustrate any change in well-being since the last interview on the well-being chart. The interview will finish by repeating the series of closed questions from the first wave. Again, a separate interview will be conducted with the most senior woman (where appropriate). The same procedures regarding the fidelity of the data will be conducted as in the first wave.

Data from the closed questions will enable the ATT to be calculated for key impact variables. Qualitative data from the life history interviews (including on changes to the intra-household division of labour and income) will be coded and analyzed in qualitative software. Some qualitative data will be quantified and compared across treatment and control groups. Visual data will also be coded using qualitative software. Again, this data will be quantified and comparisons across treatment and control groups will be drawn.

Assessing impact heterogeneity is a strength of in-depth qualitative methods such as life histories and will supplement the lack of attention paid to sub-groups in the ATT analysis. The qualitative data will allow researchers to study farmers' decision-making and intra-household processes (and the implications of these for household income and food security). This will be helpful in interpreting findings from the attribution analysis. In addition, comparisons of sub groups will be conducted (gender of household head, wealth category, landholding size) although these findings will be suggestive as reducing the sample size will reduce confidence levels.

VI Conclusion

All methodologies have limitations. Experimental design is a valuable approach (with due consideration of applicability, threats and ethics) within the spectrum available to researchers and evaluators, particularly when qualitative methods are included within the methodology. For example, Woolcock (2009: 13) views the inclusion of qualitative methods as the factor that moves a methodology from 'gold' to 'diamond' standard. Mixing methods within an experimental designs may improve the interpretation of quantitative results, avoid fundamental misunderstandings due to neglect of the context in which the intervention is taking place, foster greater engagement with evaluation communities and, more importantly, with the beneficiaries of interventions. But, as yet, there appears to be little appreciation that just because randomized experiments utilize a relatively strict positivist methodology, this does not preclude qualitative methods from taking an equal or primary role as the data measurement tool (Davis, 2010 provides a possible model). The next steps in advocating for a greater number of experimental studies that utilize a qualitative method as the primary measurement tool are to: (a) assess the implications of using qualitative methods in terms of the skills of research personnel, and institutional acceptance; and (b) conduct a detailed comparison of the interview-level strengths and shortcomings of

different measurement tools within the rubric of a randomized design. From our perspective, moving this research agenda forward chimes with Banerjee and Duflo's (2008) call for 'creative experimentalism', and may help to bridge the gap between advocates of randomized control trials and development research and evaluation communities.

Acknowledgements

The authors would like to thank Jos Vaessen and Peter Davis for comments. The usual disclaimers apply.

References

Adato, M. 2007: *Combining survey and ethnographic methods to evaluate conditional cash transfer programs*. Q-Squared Working Paper No. 40.

Banerjee, A. and **Duflo, E.** 2008: *The experimental approach to development economics*. CEPR Working Paper No. DP7037, Centre for Economic Policy and Research.

Banerjee, Abhijeet, Amsden, A.H., Bates, R.H., Bhagwati, J., Deaton, A., and **Stern, N.** et al. 2007: *Making aid work*. MIT Press.

Barahona, C. and **Levy, S.** 2003: *How to generate statistics and influence policy using participatory methods in research: Reflections on work in Malawi 1999–2002*. IDS Working Paper 212.

Cartwright, N. 2007: Are RCTs the gold standard?, *BioSocieties* 2, 11–20.

Chattopadhyay, R. and **Duflo, E.** 2004: Women as policy makers: Evidence from a randomized policy experiment in India. *Econometrica* 72, 1409–43.

Davis, P. and **Baulch, B.** 2009: *Parallel realities: Exploring poverty dynamics using mixed methods in rural Bangladesh*. Paper presented at the 'Escaping Poverty Traps: Connecting the Chronically Poor to Economic Growth' conference, Washington, DC, 26–27 February 2009.

Davis, P. 2010: *Exploring the long-term impact of development interventions within life-history narratives in rural Bangladesh*. IFPRI Discussion Paper 00991.

Deaton, A. 2009: *Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development*. The Keynes Lecture, British Academy, 9 October 2008.

Garbarino, S. and **Holland, J.** 2009: *Quantitative and qualitative methods in impact evaluation and measuring results*. Discussion Paper. University of Birmingham, Birmingham, UK.

Jones, N., Jones, H., Steer, L. and **Datta, A.** 2009: *Improving impact evaluation production and use*. ODI Working Paper 300. Overseas Development Institute, London.

Karlan, D. 2009*: Cairo evaluation clinic: Thoughts on randomized trials for evaluation of development*. Economics Department Working Paper No. 65/ Economic Growth Center Discussion Paper No. 973, Department of Economics, Yale University.

Molloy, D., Woodfield, K. and **Bacon, J.** 2002: Longitudinal qualitative research approaches in evaluation studies. A study carried out on behalf of the Department for Work and Pensions. Working Paper 7, Social Research Unit, London.

Rao, V. and **Ibáñez, A.** 2005: The social impact of social funds in Jamaica: A mixed-methods analysis of participation, targeting and collective action in community driven development. *Journal of Development Studies* 41, 788–838.

Ravallion, M. 2009: Evaluation in the practice of development. *The World Bank Research Observer* 24, 29–53.

Rodrik, D. 2009: The new development economics: We shall experiment, but how shall we learn? In Cohen, J. and Easterly, W., editors, *What works in development? Thinking big and thinking small*. Brookings Institution Press, 24–54.

Savedoff, W.D., Levine, R. and **Birdsall, N.** 2006: *When will we ever learn? Improving lives through impact evaluation. Report of the evaluation gap working group*. Center for Global Development.

White, H. 2007: Technical rigor must not take precedence over other kinds of valuable lessons. In Banerjee, A., Amsden, A.H., Bates, R.H., Bhagwati, J., Deaton, A, and Stern, N. *et al.*, editors, *Making aid work*. MIT Press, 81–91.

——— 2008: *Of probits and participation: The use of mixed methods in quantitative impact evaluation*. NONIE Working Paper No. 6, World Bank.

White, H. and **Masset, E.** 2007: The Bangladesh integrated nutrition program: Findings from an impact evaluation. *Journal of International Development* 19, 627–52.

Woolcock, M. 2009: *Towards a plurality of methods in project evaluation: A contextualised approach to understanding impact trajectories and efficacy*. BWPI Working Paper 73.